

Hacia un método computacional para detectar léxico contrastivo

Damián Eliel Aleman

9 de noviembre de 2017

- 1 Introducción
- 2 Trabajo Previo
- 3 Datos: Extracción y procesamiento
- 4 Métricas para detectar léxico contrastivo
- 5 Análisis de las palabras contrastivas encontradas
 - Caracterización de las palabras identificadas como contrastivas
 - Validación estadística
- 6 Conclusiones y trabajo futuro

Qué es una palabra contrastiva

Se dice que una palabra es contrastiva cuando la frecuencia de uso en distintas regiones es muy diferente.

Ejemplos palabras contrastivas Argentina - España

- “che”
- “metegol”

Ejemplos palabras contrastivas dentro de Argentina

- “gurisada”

¿Para qué sirve conocer las palabras contrastivas?

Motivación de un método computacional para detectar léxico contrastivo

¿Cómo se conocían las palabras contrastivas? Mediante encuestas.

- Es costoso de realizar
- Muy difícil de hacer de forma balanceada en distintas regiones de un país o de un continente.
- Más difícil es encuestar a una gran cantidad de personas.
- Se basan en el conocimiento *a priori*

- 1 Introducción
- 2 Trabajo Previo
- 3 Datos: Extracción y procesamiento
- 4 Métricas para detectar léxico contrastivo
- 5 Análisis de las palabras contrastivas encontradas
 - Caracterización de las palabras identificadas como contrastivas
 - Validación estadística
- 6 Conclusiones y trabajo futuro

Web como un corpus

Kilgariff comentaba la riqueza de la Web para acceder a una información, antes impensada. Principalmente se hacían estudios estimando la frecuencia de una palabra viendo la cantidad de resultados de las consultas en un motor de búsqueda.

Ventajas:

- Gratuito
- Gran cantidad de datos
- Disponible e inmediato

Utilidades:

- Corrección de ortografía
- Traducción de frases
- Estimar el tamaño de la Web

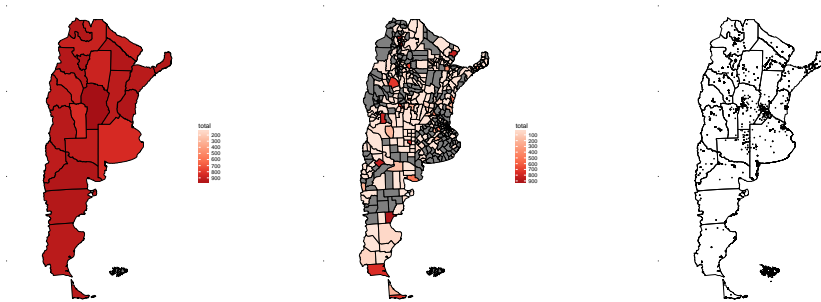
Cita

“La web es un corpus sucio, pero el uso esperado es mucho más frecuente que lo que puede considerarse como ruido.” [KG03, p. 342]

- 1 Introducción
- 2 Trabajo Previo
- 3 Datos: Extracción y procesamiento**
- 4 Métricas para detectar léxico contrastivo
- 5 Análisis de las palabras contrastivas encontradas
 - Caracterización de las palabras identificadas como contrastivas
 - Validación estadística
- 6 Conclusiones y trabajo futuro

Búsquedas geolocalizadas

La búsqueda geolocalizada es una herramienta que nos da la posibilidad de obtener tuits generados en un área geográfica particular. Para esto, primero intenta buscar tuits cuyas coordenadas sean las buscadas.



Se realizaron búsquedas por cada provincia con centro en las coordenadas de los departamentos de la misma y con un radio de 20 millas. Nos quedamos con los usuarios que tienen como campo *location* al menos uno de los nombres de las ciudades de la provincia.

Temario

- 1 Introducción
- 2 Trabajo Previo
- 3 Datos: Extracción y procesamiento
- 4 Métricas para detectar léxico contrastivo**
- 5 Análisis de las palabras contrastivas encontradas
 - Caracterización de las palabras identificadas como contrastivas
 - Validación estadística
- 6 Conclusiones y trabajo futuro

Temario

- 1 Introducción
- 2 Trabajo Previo
- 3 Datos: Extracción y procesamiento
- 4 Métricas para detectar léxico contrastivo
- 5 Análisis de las palabras contrastivas encontradas
 - Caracterización de las palabras identificadas como contrastivas
 - Validación estadística
- 6 Conclusiones y trabajo futuro

- 1 Introducción
- 2 Trabajo Previo
- 3 Datos: Extracción y procesamiento
- 4 Métricas para detectar léxico contrastivo
- 5 Análisis de las palabras contrastivas encontradas
 - Caracterización de las palabras identificadas como contrastivas
 - Validación estadística
- 6 Conclusiones y trabajo futuro



Adam Kilgarriff and Gregory Grefenstette.

Introduction to the special issue on the web as corpus.

Computational linguistics, 29(3):333–347, 2003.